Queue Waiting Time Estimation Using Person Re-identification by Upper Body

Timur Mamedov[1,2], Denis Kuplyakov[1,2] and Anton Konushin[1,3]

[1]Lomonosov Moscow State University, 1, Leninskie Gory, Moscow, 119991, Moscow, Russia
[2]Video Analysis Technologies, 7, Sculptora Mukhina, 119634, Moscow, Russia
[3]NRU Higher School of Economics, 11, Pokrovsky Bulvar, 109028, Moscow, Russia

Abstract

In this paper, we propose a new approach to estimating waiting time in queue based on object tracking and person re-identification by upper body. The task we are considering is practically important in video analysis, because this data can be used for predictive analytics and improvement of customer services. The main idea of the proposed method is to use upper body detections instead of full body detections. This decision is due to the following fact: in queues, the upper bodies are more visible. Using re-identification allows us to perform video analytics on sparse frames and thereby increase the computational efficiency of the estimation algorithm. Also in this work, we introduce a novel upper body regression by head, upper body random size augmentation to improve re-identification performance in real-world scenarios and a method for calculating metrics for queue waiting time estimation algorithms. Our experimental evaluation showed that the proposed algorithm has a high accuracy of queue waiting time estimation.

Keywords

Computer Vision, Video Analytics, Re-identification, Object Tracking, Queue Analytics

1. Introduction

The task of estimating waiting time in queue is practically an important task, since algorithms for solving this problem are used in commercial products to improve customer services. There are many solutions to this problem, one of them is object tracking. The task of object tracking is to create tracks for each person and every track unambiguously corresponds to a person. It marks this particular person locations in all frames in which person is visible.

We use appearance embeddings obtained using the re-identification algorithm (next, we will call it as Re-ID appearance embeddings) to bind detections on new frames with existing tracks or to create new tracks. Using re-identification allows us to perform video analytics on sparse frames and thereby increase the computational efficiency of the estimation algorithm. This fact is very important for real-world tasks, because in practice we have strict limits on computing resources. Also, in this paper, we use upper body detections for re-identification algorithm

GraphiCon 2021: 31st International Conference on Computer Graphics and Vision, September 27–30, 2021, Nizhny Novgorod, Russia

timur.mamedov@graphics.cs.msu.ru (T. Mamedov); denis.kuplyakov@graphics.cs.msu.ru (D. Kuplyakov); anton.konushin@graphics.cs.msu.ru (A. Konushin)
0000-0001-6554-7988 (T. Mamedov); 0000-0002-2957-3297 (D. Kuplyakov); 0000-0002-6152-0021 (A. Konushin)

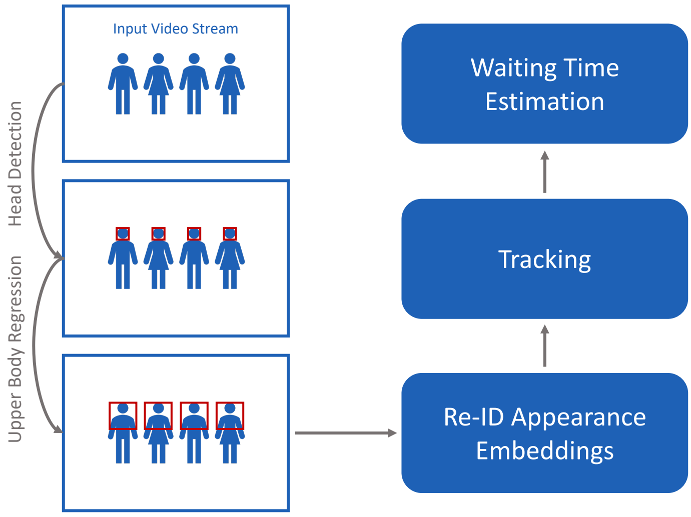


Figure 1: Scheme of the proposed method for estimating waiting time in queue.

instead of full body detections, because in queues, the upper bodies are more visible.

In this work, we proposed a fully automatic algorithm for estimating waiting time in queue. The input of the algorithm is a video stream $\{F_i\}_{i=1}$ of frames captured by single stationary camera, the coordinates of the region of interest (ROI) and t — fragment length in seconds. The output of the algorithm is a set $\{T_i\}_{i=1}$ of maximum waiting time every t seconds.

Summarize, in this article we offer the following ideas:

- novel upper body regression by head;
- new upper body random size augmentation;
- neural network algorithm for person re-identification by upper body detections;
- object tracking algorithm based on Re-ID appearance embeddings for creating tracks;
- new method for calculating metrics for queue waiting time estimation algorithms.

2. Related Work

Most of the existing methods for queue monitoring use idea of counting people who have crossed a special signal line, which located at the entrance and exit of the queue. For this purpose, special equipment is often used, which is installed at the entrance and exit of the queue, for example, infrared sensors in [1]. However, there are methods based on computer vision techniques. For example, the work [2] uses tracking to count people who have crossed the signal line. In this paper we also use object tracking algorithm, but we abandoned the idea of signal lines. To estimate the waiting time in the queue, we focus on the lifetime of the tracks.

There are a lot of types of object tracking algorithms, most of them are based on tracking-by-detection. There are several approaches to detecting objects on the frame for further tracking. For example, detection of body [3, 4, 5], detection of head [2, 6] and using of key points. Neural networks methods also used in object tracking task, for example, in [7] detector is used to regress detection on the current frame by detection on previous frame.

As mentioned earlier, we use person Re-ID appearance embeddings to bind detections to the tracks. Re-identification algorithms have been particularly developed with the growing popularity of neural network methods. Occlusions — is the one of the main problems in re-identification task. To solve the problem of occlusions, binary object masks [8, 9] and semantic segmentation [10] are very often used. Loss functions [11] and neural network architectures [12] also play a huge role in re-identification.

3. Proposed Method

The fig. 1 shows the general scheme of the proposed method. Our method consist of five main steps:

- **Head Detection** — at this step, the head detector is used to search for the heads of all people on sparse frames;
- **Upper Body Regression** — at this step, our proposed upper body regression by heads is used to regress the upper bodies for each head found on sparse frames;
- **Re-ID Appearance Embeddings** — at this step, all the regressed upper bodies are fed to the input of the neural network for re-identification, which generates embeddings;
- **Tracking** — at this step, Re-ID appearance embeddings are used to bind detections to the tracks or create new tracks;
- **Waiting Time Estimation** — after tracking we get set of tracks $\{Tr_i\}_{i=1}$, $Tr_i = \{f_j, d_j\}_{j=1}$, where f_j, d_j — frame and coordinates of the bounding box, respectively. Using the set of tracks, we can calculate waiting time in the queue as the maximum lifetime (by the lifetime $L(tr)$ of the track tr, we mean the duration of its existence in a video sequence) of one of the tracks from this set.

Below are detailed descriptions of the first four steps of the proposed algorithm.

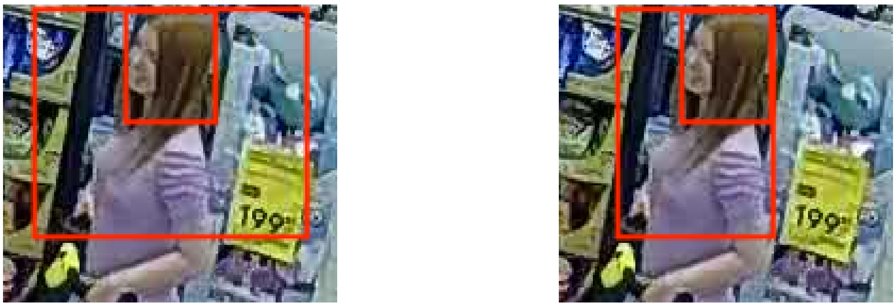
3.1. Head Detection

In this paper, we use the head detector based on SSD [13] with ResNet50 [14] as the backbone. We choose this detector, because it is fast and has acceptable quality for the detection task. The detector were trained on CrowdHuman public dataset [15] and on the dataset collected by Video Analysis Technologies. Experimental evaluation showed AUC of 0.66 on the test part of CrowdHuman.

Next, head detections are used to regress the upper parts of people's bodies.

3.2. Upper Body Regression

The main idea of the proposed method is to use upper body detections instead of full body detections for re-identification algorithm. This decision is due to the following fact: in queues, the upper bodies are more visible. Also, the upper body contains sufficient visual information



(a) Heuristic

(b) Neural Network Regression

Figure 2: Comparison of heuristics and upper body neural network regression.

for re-identification compared to the heads. Our upper body regression approach consist of two steps:

1. using heuristics, we find the approximate position of the upper body;
2. using neural network regression, we clarify the position of the upper body.

The First Step In our heuristics we are using the following fact from anatomy: the width of a person's body is on average three times the width of a person's head, and the height of a person's body is on average eight times the height of a person's head.

Under the upper part of the human body in this work, we understand the part of the body that is equal in height to two heights of the human head, then:

$$upperBody_{left} = head_{left} - head_{width}, \tag{1}$$

$$upperBody_{top} = head_{top}, \tag{2}$$

$$upperBody_{width} = 3 \cdot head_{width}, \tag{3}$$

$$upperBody_{height} = 2 \cdot head_{height}, \tag{4}$$

where $(head_{left},\ head_{top},\ head_{width},\ head_{height})$ — the coordinates of the head bounding box and $(upperBody_{left},\ upperBody_{top},\ upperBody_{width},\ upperBody_{height})$ — the coordinates of the upper body bounding box.

The Second Step At the second step we specify the exact position of the upper body bounding box, obtained in the previous step. To do this, we developed neural network to regress two coordinates: the left and right extreme points of the upper part of the human body ($upperBody_{regLeft}$ and $upperBody_{regRight}$, respectively). This coordinate refinement is especially important if the person on the frame is standing in profile (see fig. 2). The fig. 3 shows the architecture of the proposed neural network for upper body regression. Our regression neural network has a simple architecture, because we need our algorithm for estimating waiting time in queue to work in real time. Despite the simple architecture, the proposed neural network has a good regression quality.

As a result, we have the final coordinates of the upper body bounding box, which is further used in re-identification algorithm: $(upperBody_{regLeft},\ upperBody_{top},\ upperBody_{regWidth},\ upperBody_{height})$, where $upperBody_{regWidth} = upperBody_{regRight} - upperBody_{regLeft}$.

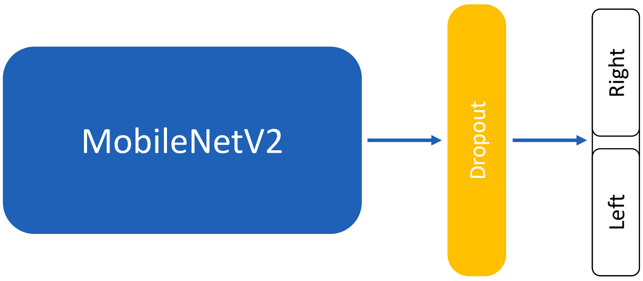


Figure 3: The architecture of the proposed neural network for upper body regression. We use MobileNetV2 [16] as the backbone of our neural network.

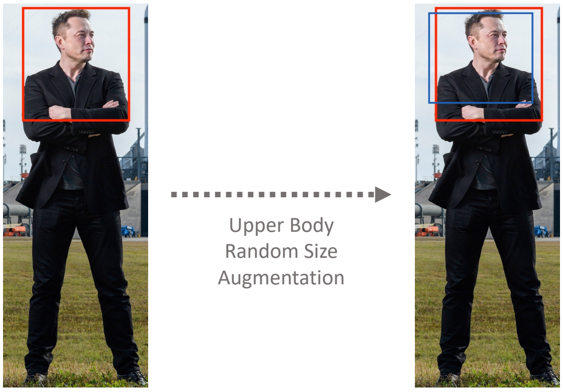


Figure 4: An example of the proposed Upper Body Random Size Augmentation. Red bbox is a detection, blue bbox is the modified red bbox using the suggested augmentation.

3.3. Re-ID Appearance Embeddings

Re-identification Neural Network We use the solution from [12] as the baseline for person re-identification, because this algorithm has high performance in the re-identification task and it works quickly, which is important in the problem we are solving. We applied two modifications to the baseline:

- ResNet50 was replaced by Res2Net50 [17], because the second neural network has a higher quality on the classification task, but at the same time it is slightly slower in speed than ResNet50;
- we use FIDI Loss [11] instead of Triplet Loss and Center Loss [18] in the baseline. This loss function greatly penalizes small differences between images, which is very important for the re-identification task.

Upper Body Random Size Augmentation Consider, for example, the case of full body re-identification. Usually, when training a neural network, images are used for re-identification, in which people are represented in full growth. However, in real-world scenarios, due to occlusions (or detector mistakes), there may be cases where the detector finds only a part of the body, for example, only the upper part of the body. The use of such detections for re-identification can seriously reduce the quality of person re-identification, since the neural network has previously “seen” people only in full growth.

The same is possible for person re-identification by upper body. To solve this problem, we introduce a novel upper body random size augmentation to improve re-identification performance in real-world scenarios. The main idea of the proposed augmentation is to randomly change the boundaries of the upper body bounding box (see fig. 4).

Let $(\mathit{left}_1,\ \mathit{top}_1,\ \mathit{width}_1,\ \mathit{height}_1)$ — the coordinates of the upper body bbox obtained by the regressor, $(\mathit{left}_2,\ \mathit{top}_2,\ \mathit{width}_2,\ \mathit{height}_2)$ — the coordinates of the upper body bbox after using the proposed augmentation, *width* and *height* — width and height of the full body bbox, then:

$$\mathit{left}_2 = \max(0,\ \mathit{left}_1 + \alpha_1 \cdot random(-\mathit{width}_1,\ \mathit{width}_1)), \tag{5}$$

$$\mathit{top}_2 = \max(0,\ \mathit{top}_1 + \alpha_2 \cdot random(-\mathit{height}_1,\ \mathit{height}_1)), \tag{6}$$

$$\mathit{width}_2 = \min(\mathit{width},\ \mathit{left}_1 + \mathit{width}_1 + \alpha_3 \cdot random(-\mathit{width}_1,\ \mathit{width}_1)) - \mathit{left}_2, \tag{7}$$

$$\mathit{height}_2 = \min(\mathit{height},\ \mathit{top}_1 + \mathit{height}_1 + \alpha_4 \cdot random(-\mathit{height}_1,\ \mathit{height}_1)) - \mathit{top}_2, \tag{8}$$

where $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ — pre-selected constants, in this paper we use $\alpha_1 = \alpha_3 = \alpha_4 = 0.25$ and $\alpha_2 = 0.05$, and $random(a, b)$ is a function that returns a random integer value $x \in [a, b]$.

3.4. Tracking

We use the solution from [3] as the baseline for object tracking. Baseline works in online mode and use Hungarian algorithm [19] to match detections. In this work, we use appearance embeddings obtained using the upper body re-identification algorithm to bind detections on new frames with existing tracks or to create new tracks. For each track we store 100 most recent re-identification descriptors that were used to bind new detections to the track.

Let $\{D_i\}_{i=1}^{N}$ is a set of re-identification descriptors for detections on new frame and $\{Tr_i\}_{i=1}^{M}$ is a set of lists of re-identification descriptors for all existing tracks. Let’s set the cost matrix $CM \in \mathbb{R}^{N\times M}$ as follows:

$$c_{i,j} = \max_{tr_{j_k} \in\, tr_j} CosineSimilarity(d_i, tr_{j_k}),\ \text{where } tr_j \in \{Tr_i\}_{i=1}^{M},\ i \in [1, N],\ j \in [1, M], \tag{9}$$

$$CM[i, j] = \begin{cases} c_{i,j}, & \textbf{if } c_{i,j} \geqslant thr \\ 0, & \textbf{else} \end{cases},\ i \in [1, N],\ j \in [1, M], \tag{10}$$

where thr is the threshold for the cosine similarity, in this article we use $thr = 0.4$. Next, for cost matrix CM the assignment problem is solved using the Hungarian algorithm to maximize cost.

4. Experiments

4.1. Datasets

Upper Body Regression To train our regression neural network (see section 3.2) we used a modified CrowdHuman public dataset [15]. Our modifications are as follows:

1. using the head detector (see section 3.1), head detections were obtained for each image;
2. using the Detectron2 [20], body detections and segmentation masks were obtained for each image;
3. using a heuristic from [2], the head and body detections were mapped for each image;
4. using our heuristic (see section 3.2), the approximate positions of the upper bodies were found for each image. This data was used to make crops for training our upper body regression neural network;
5. using segmentation masks for the body detections corresponding to the head detections (according to the detection of the upper body detections), the left and right extreme points of the upper part of the human body were found for each image. This data was used to train our upper body regression neural network.

Re-identification To train our re-identification neural network we used a modified MSMT17 [21] public dataset that combined the training and test parts (we named the resulting dataset MSMT17 Merged). Our modification is as follows: we used our upper body regression neural network to find the upper body detection for each image. The obtained upper body detections were used to train person re-identification by upper body.

Waiting Time Estimation Algorithm To test the entire algorithm proposed in this paper we need datasets captured by static camera with head tracks markup. Also, the video sequences should be long enough to evaluate the quality of the estimating waiting time in queue. We used 6 videos from the collection of the Video Analysis Technologies company. These videos are obtained from real security cameras in stores and other public places where queues are possible, and allow us to bring testing closer to real-world scenarios. In addition, we have developed an effective procedure for marking up long real videos. The table 1 provides detailed information about each test video.

Table 1
Videos that we used to test the Waiting Time Estimation Algorithm.

Video Name	Duration	Format
Queue/1	00:35:00	1920×1080, 25 FPS
Queue/2	00:32:00	704×576, 15 FPS
Queue/3	00:18:00	2688×1520, 24 FPS
Queue/4	00:30:00	1280×720, 60 FPS
Queue/5	01:24:00	1920×1080, 25 FPS
Queue/6	00:30:00	1280×720, 60 FPS

4.2. Metrics

In this paper, we propose a new method for calculating metrics for queue waiting time estimation algorithms. Our method is as follows:

1. we divide videos to segments of equal length t seconds (with the possible exception of the last segment). In this paper, $t = 300$ seconds;

2. for each segment we calculate $GT_{seg} = \max_{tr \in Tr_{GT_{seg}}} L(tr)$ and $Pred_{seg} = \max_{tr \in Tr_{Pred_{seg}}} L(tr)$, where $Tr_{GT_{seg}}$ and $Tr_{Pred_{seg}}$ are sets of the ground truth and predicted tracks on the segment, respectively. $L(tr)$ is the lifetime of the track tr on the segment. That is, for each segment in the ground truth and the results of the algorithm, a track with the maximum lifetime on the segment is sought. **Important:** track lifetimes are end-to-end, meaning they are shared across the entire video, not separate for each segment;
3. for each segment we calculate the absolute and relative error:

$$AE_{seg} = |Pred_{seg} - GT_{seg}|, \tag{11}$$

$$RE_{seg} = \frac{AE_{seg}}{GT_{seg}}; \tag{12}$$

4. for the entire video we calculate the mean absolute error and mean relative error in percentage:

$$MAE = \frac{\sum_{seg=1}^{N} AE_{seg}}{N}, \tag{13}$$

$$MRE = \frac{\sum_{seg=1}^{N} RE_{seg}}{N} \cdot 100\%, \tag{14}$$

where N is the number of segments.

4.3. Experimental Results

In this section, we present experimental results for the entire proposed algorithm. We consider 4 types of experiments:

- **Detectron2 Det.** — in this experiments, we used person re-identification by full body and detections from Detectron2 [20];
- **Upper Body Heuristic Det.** — in this experiments, we used person re-identification by upper body and detections from our heuristic (see section 3.2);
- **Upper Body Reg. Det.** — in this experiments, we used person re-identification by upper body and detections from proposed neural network for upper body regression (see section 3.2);
- **Upper Body Reg. Det. and UBRSA** — in this experiments, we used person re-identification by upper body with a new Upper Body Random Size Augmentation (see section 3.3) and detections from proposed neural network for upper body regression (see section 3.2).

The table 2 provides detailed information on the results of each of the above experiments.

The experimental evaluation shows that our hypothesis that the upper parts of the body are seen better than the full bodies in the queues is correct. In addition, huge error in experiments with re-identification by full body and detections from Detectron2 related to the fact described in the section 3.3.

A comparison of the results of experiments with heuristic upper body detections and experiments with neural network upper body detections shows that the proposed regression

Table 2
Experimental results for the entire proposed algorithm.

Algorithm	**Frames every 5 sec.**		**Frames every 10 sec.**	
	MAE (↓), sec.	MRE (↓), %	MAE (↓), sec.	MRE (↓), %
Detectron2 Det.	247.5	34.3	253.8	41.7
Upper Body Heuristic Det.	51.2	15.0	40.4	12.8
Upper Body Reg. Det.	33.9	8.8	32.1	**10.7**
Upper Body Reg. Det. and UBRSA	**23.1**	**6.3**	**29.4**	11.9

strategy has a positive effect on the quality of estimating waiting time. This is because heuristic detections often have wide bboxes that either get a lot of background or other people into them (see fig. 2). And this entails a re-identification error.

The experimental evaluation shows that the proposed Upper Body Random Size Augmentation for person re-identification also has a positive effect on the quality of estimating waiting time, since this augmentation solves the problem described in the section 3.3. Solving this problem allows us to reduce the number of broken tracks and, as a result, improve the quality of object tracking.

5. Conclusion

We have proposed the algorithm to estimating waiting time in queue based on object tracking and person re-identification by upper body. Using re-identification allows us to perform video analytics on sparse frames and thereby increase the computational efficiency of the estimation algorithm. In this work, we have introduced a method for calculating metrics for queue waiting time estimation algorithms. In addition, we have proposed a novel upper body regression by head, upper body random size augmentation to improve re-identification performance in real-world scenarios, which improved quality of the algorithm.

References

[1] U. A. Gimba, C. D. Okoronkwo, M. Yusuf, A. S. Musa, M. S. Ali, Queue monitoring system for bank, Dutse Journal of Pure and Applied Sciences (DUJOPAS) 6 (2020) 269–276.

[2] D. Kuplyakov, Y. Geraskin, T. Mamedov, A. Konushin, A distributed tracking algorithm for counting people in video by head detection, in: Proceedings of the 30th International Conference on Computer Graphics and Machine Vision, volume 2744 of *CEUR Workshop Proceedings*, M. Jeusfeld c/o Redaktion Sun SITE, Informatik V, RWTH Aachen, 2020, pp. 1–12. doi:10.51130/graphicon-2020-2-3-26.

[3] A. Bewley, Z. Ge, L. Ott, F. Ramos, B. Upcroft, Simple online and realtime tracking, 2016 IEEE International Conference on Image Processing (ICIP) (2016). URL: http://dx.doi.org/10.1109/ICIP.2016.7533003. doi:10.1109/icip.2016.7533003.

[4] N. Wojke, A. Bewley, D. Paulus, Simple online and realtime tracking with a deep association metric, 2017. arXiv:1703.07402.

[5] F. Yu, W. Li, Q. Li, Y. Liu, X. Shi, J. Yan, Poi: Multiple object tracking with high performance detection and appearance feature, in: European Conference on Computer Vision, Springer, 2016, pp. 36–42.

[6] D. Kuplyakov, E. Shalnov, A. Konushin, Further improvement on an mcmc-based video tracking algorithm, in: Proceedings of the 26th International Conference on Computer Graphics and Vision GraphiCon'2016, GraphiCon, 2016, p. 440–444.

[7] P. Bergmann, T. Meinhardt, L. Leal-Taixé, Tracking Without Bells and Whistles, in: The IEEE International Conference on Computer Vision (ICCV), 2019.

[8] C. Song, Y. Huang, W. Ouyang, L. Wang, Mask-guided contrastive attention model for person re-identification, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 1179–1188. doi:10.1109/CVPR.2018.00129.

[9] H. Cai, Z. Wang, J. Cheng, Multi-scale body-part mask guided attention for person re-identification, volume abs/1904.11041, 2019. URL: http://arxiv.org/abs/1904.11041. arXiv:1904.11041.

[10] M. M. Kalayeh, E. Basaran, M. Gokmen, M. E. Kamasak, M. Shah, Human semantic parsing for person re-identification, volume abs/1804.00216, 2018. URL: http://arxiv.org/abs/1804.00216. arXiv:1804.00216.

[11] C. Yan, G. Pang, X. Bai, J. Zhou, L. Gu, Beyond triplet loss: Person re-identification with fine-grained difference-aware pairwise loss, 2020. arXiv:2009.10295.

[12] H. Luo, Y. Gu, X. Liao, S. Lai, W. Jiang, Bag of tricks and a strong baseline for deep person re-identification, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2019.

[13] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A. C. Berg, Ssd: Single shot multibox detector, Lecture Notes in Computer Science (2016) 21–37. URL: http://dx.doi.org/10.1007/978-3-319-46448-0_2. doi:10.1007/978-3-319-46448-0_2.

[14] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2015. arXiv:1512.03385.

[15] S. Shao, Z. Zhao, B. Li, T. Xiao, G. Yu, X. Zhang, J. Sun, Crowdhuman: A benchmark for detecting human in a crowd, 2018. arXiv:1805.00123.

[16] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, Mobilenetv2: Inverted residuals and linear bottlenecks, 2019. arXiv:1801.04381.

[17] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, P. Torr, Res2net: A new multi-scale backbone architecture, volume 43, Institute of Electrical and Electronics Engineers (IEEE), 2021, p. 652–662. URL: http://dx.doi.org/10.1109/TPAMI.2019.2938758. doi:10.1109/tpami.2019.2938758.

[18] Y. Wen, K. Zhang, Z. Li, Y. Qiao, A discriminative feature learning approach for deep face recognition, in: B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.), Computer Vision – ECCV 2016, Springer International Publishing, Cham, 2016, pp. 499–515.

[19] H. W. Kuhn, The hungarian method for the assignment problem, Naval Research Logistics Quarterly 2 (1955) 83–97. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/nav.3800020109. doi:https://doi.org/10.1002/nav.3800020109. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/nav.3800020109.

[20] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, R. Girshick, Detectron2, https://github.com/facebookresearch/detectron2, 2019.

[21] L. Wei, S. Zhang, W. Gao, Q. Tian, Person transfer gan to bridge domain gap for person re-identification, 2018. arXiv:1711.08565.